A comprehensive catalog of predicted functional upstream open reading frames.

Patrick McGillivray[1], Russell Ault[1], Mayur Pawashe[1], Rob Kitchen[2], Suganthi Balasubramanian[1,2,3*], Mark Gerstein[1,2,4*]

[1]Molecular Biophysics and Biochemistry Department, Yale University, New Haven 06520, CT, USA
[2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520 USA
[3]Regeneron Genetics Center, Tarrytown, NY 10591, USA
[4]Department of Computer Science, Yale University, New Haven 06520, CT, USA

* Corresponding authors: suganthi.bala@regeneron.com, pi@gersteinlab.org
$ Current address

Abstract:

Upstream open reading frames (uORFs) latent in mRNA transcripts are thought to modify translation of coding sequences by altering local ribosome activity. Not all uORFs are thought to be active in such a process.

To estimate the impact of uORFs on regulation of translation, we first circumscribed the universe of all uORFs based on coding gene sequence, and identified over one million unique uORFs. To determine which of these uORFs are likely biologically relevant, we built a classifier using 89 attributes of uORFs labeled as active in experiment. Our classifier allowed us to extrapolate to a catalog of uORFs that are likely active from the universe of all uORFs.

This is a substantially larger catalog of uORFs than has previously been associated with active function. Our ranked list of likely active uORFs allows researchers to test their hypotheses regarding the role of upstream open reading frames in health and disease. We demonstrate several examples of biological relevance through application of our catalog.

Introduction:

Upstream open reading frames (uORFs) consist of a start codon in the 5' untranslated region of a gene (UTR) and an associated stop codon appearing before the stop codon of the main coding DNA sequence (CDS). An uORF may begin and end before the main gene coding sequence. Alternatively, if the upstream reading frame is out of frame with the CDS, it may overlap with the CDS [Fig. 1a]. uORFs are latent in mRNA transcripts and may undergo partial or complete translation.

An initial survey of the human genome identified uORFs contained in approximately 10% of mRNA transcripts [1]. More recent analyses identify uORFs in association with nearly half of all mRNA transcripts [2]. The discovery that many translated uORFs utilize near-cognate start codons to the canonical ATG start codon has broadened estimates of uORF prevalence further [3–6].

Presence of functional uORFs is generally thought to suppress translation of downstream genes [7–12] [Fig. 1b]. Proposed molecular mechanisms for modification of CDS translation by uORFs are numerous. These include *translation reinitiation* -- the uORF and CDS are translated by the same ribosome in series -- *leaky-scanning* -- ribosome recognition of an uORF and subsequent CDS translation, without uORF translation -- and *ribosome-stalling* -- decreased translation of the CDS, due to ribosome retention at the upstream uORF [3,13,14]. Differential translation of multiple protein products may occur in consequence to an uORF [15]. It is also possible for an uORF to function as a short open reading frame, encoding a functional peptide [16–19]. uORF function is not necessarily constant -- uORFs may display differential function in stressed cells, compared with non-stressed controls [20–25].

Study of uORF translation and function was historically limited to the experimental evaluation of individual uORFs [7,26]. Genome-scale ribosome profiling studies have allowed for the identification of large populations of uORFs known to undergo translation [4,27,28]. This mapping of translation initiation is sufficient for association between ribosomes and particular start codons and reading frames [29–31].

We proceed on the assumption that the total universe of active uORFs is much larger than that identified through ribosome profiling experiments. In other words, we assume that ribosome profiling experiments have high specificity in identifying functional uORFs with a high false-negative rate [Fig. 1c]. Ribosome profiling experiments follow a challenging technical procedure, and it is uncertain whether all potentially active uORFs are measurable in a given sample [Fig. 1d]. This is consistent with a high false-negative rate. Other researchers have implicitly endorsed this hidden assumption, when predicting translated uORFs in *Saccharomyces cerevisiae* and *Arabidopsis thaliana,* on the basis of DNA sequence and ribosome profiling data [32,33]. A similar assumption is the basis for using patterns of ribosome profiling occupancy to maximize the number of inferred translation products in humans [34,35].

For our investigation of the prevalence of active uORFs in humans, we began with a genome wide scan, searching for uORFs associated with protein coding genes listed in the GENCODE genome annotation [36]. All possible uORFs beginning with ATG, or a single nucleotide variant of ATG, were identified. This scan yields a universe of all possible uORFs numbering nearly 1.3 million.

uORFs in this large set were classified as active according to similarity to uORFs occupied in ribosome profiling experiments. This classification was accomplished using a Naïve-Bayes classifier, trained on 89 uORF attributes. We validated our predicted uORFs using a cross-validation method where two ribosome profiling experiments are used to predict the uORFs translated in a third experiment. We also validated our predictions by examining how gene level protein expression and local ribosome activity correlate with genetic variants that alter uORFs in 46 individuals.

The 1000 Genomes Project's database of human variation [37] and the NHGRI-EBI GWAS catalog [38] were used to provide a baseline for the functional consequence of our predicted active uORFs. The predictions we generated were also used to measure the functional impact of somatic mutations affecting uORFs, in tissue-matched tumor samples [39].

We provide a resource of predicted active uORFs for other scientists to use in their effort to understand uORF function in health and disease.

Methods:

Extracting uORFs from GENCODE:

uORFs were identified through genome-wide search, performed on v19 of GENCODE's human genome annotation [36]. uORFs were defined as a start codon within the 5'UTR and a downstream stop codon before the end of the CDS. All three possible reading frames were examined. ATG and near cognate start codons were included in this search [ATG, TTG, GTG, CTG, AAG, AGG, ACG, ATA, ATT, ATC].

Ribosome profiling experiments as a reference set:

The ribosome profiling experiments of Lee et al. (2012), Fritsch et al. (2012) and Gao et al. (2014) were used to obtain an experimentally validated set of translated upstream open reading frames. These studies identify translation initiation sites (TIS) through treatment of human cell lines with antibiotic translation inhibitors. These treatments reliably halt ribosomes in predictable proximity to the start codon (12-13 nucleotides downstream). As such, these experiments provide high resolution information about translation initiation sites in the human genome.

Read alignments and identification of translation initiation sites were provided by these three groups of researchers. The cell lines, treatment protocols, and TIS identification mechanisms employed by each of these three research groups are summarized in *Supplemental Table 1*.

Literature review of translated human uORFs:

In addition to ribosome profiling studies, confirmed translated uORFs were obtained from the biomedical literature [7,40,41]. uORFs studied in humans that displayed functionality -- demonstrated regulation of the CDS product -- were added to the set of positive uORFs. In total, 33 uORFs, associated with 33 separate genes, were included from this literature review.

Cleansing the data set, by removal of N-terminal extensions and aTISs, and isolation of unique transcript IDs:

N-terminal extensions of the CDS sequence may retain some functional activity of the primary gene protein product, and were removed from the data set. Any uORF start codon annotated as an alternative translation initiation site (aTIS) for the CDS was also removed from the data set.

Multiple transcripts may share the same uORF. In order to avoid over-counting, only one transcript ID is attributed to a given uORF. This selection was made randomly, from among transcripts with identical chromosomal coordinates.

1-voted, 2-voted, and unlabeled data sets:

uORFs were divided into three separate sets according to their experimental translation status:

2-voted: uORFs identified as translated in two or more ribosome profiling experiments, or through literature review.
1-voted: uORFs identified as translated in only one ribosome profiling experiment.

Unlabeled: uORFs that were not identified as translated in any ribosome profiling experiment, or through literature review.

Estimating the total population of active uORFs:

Based on observed overlap among ribosome profiling experiments, an estimate for the total number of active uORFs was made using methods from population biology. Ribosome profiling experiments are treated as independent population samplings, and the Schnabel equation (Eq. 1) or Schumacher and Eschmeyer equation (Eq. 2) provide a population size estimate:

$$\boldsymbol{\widehat{N} = \frac{\sum_{t=1}^{S}(C_t\, M_t)}{\sum_{t=1}^{S} R_t}} \tag{1}$$

$$\boldsymbol{\widehat{N} = \frac{\sum_{t=1}^{S}(C_t\, M_t^2)}{\sum_{t=1}^{S} R_t\, M_t}} \tag{2}$$

Where $\widehat{N}$ is an estimate of the number of individuals in a population, given a series of *S* samplings taken at times t ∈ {1…S}, with C_t the number of individuals 'captured' in a sample, M_t the cumulative number of marked individuals prior to sampling at time *t*, and R_t the number of marked individuals 'recaptured' at sampling *t*.

Extraction of attributes associated with uORFs:

Feature data was extracted for each uORF. Features were chosen to cover a broad range of categories of data, including features associated with uORF structure, uORF evolutionary conservation, and genomic context. 89 features were used. A complete listing of these features, including details relating to the extraction and calculation of each feature, is included in *Supplemental Methods*.

Feature discretization:

The minimum description length principle (MDLP) algorithm was used to discretize each of our chosen attributes [42]. The MDLP algorithm minimizes information lost through discretization. MDLP discretization was implemented using the 'discretization' package available for R (http://cran.r-project.org/web/packages/discretization/index.html).

Prioritization of feature data:

The distribution for each feature was compared between positive and unlabeled uORFs using the Kolmogorov-Smirnov (KS) statistic. A greater KS statistic suggests greater ability of that attribute to distinguish between positive and unlabeled features.

Classifying uORFs according to attributes:

We determined that attributes of an uORF were consistent with an active uORF according to a Naive-Bayes machine learning algorithm applied to positive and unlabeled examples [43]:

$$\boldsymbol{P_{pos}\prod_{i=1}^{N} p(A_i|pos) = p_{pos}} \tag{3}$$

$$P_{neg}\prod_{i=1}^{N} p(A_i|unl) = p_{neg} \tag{4}$$

Where:

$$P_{neg} + P_{pos} = 1 \tag{5}$$

P_{pos} is the prior probability associated with positive uORFs. P_{pos} was chosen as the F1 score maximizing value (0.61). $p(A_i|pos)$, and $p(A_i|unl)$ represent the frequency of that attribute value among the positive and unlabeled sets respectively. p_{pos} represents the probability the uORF is positive. $pneg$ represents the probability the uORF is negative. We label an uORF as positive or negative according to the greater value between p_{pos} and p_{neg}. We note likely violation of the feature independence requirement of Naive-Bayes. However, empirical and theoretical study has demonstrated optimal classification performance, even where feature independence does not hold [44,45].

Model validation:

Our model was serially trained on two of three ribosome profiling data sets, using the trained model to extract the third withheld ribosome profiling data set from among the unlabeled examples. The success of differentially trained models in this cross-validation was evaluated using ROC curves, with area under the curve (AUC) calculated for each curve.

As biologic validation of our predicted uORFs, we examined the effect of alteration of a predicted active uORF's start codon on gene protein levels and local ribosome occupancy. Protein levels and local ribosome quantitative trait loci (cis-rQTL) for 46 individuals were obtained from the ribosome profiling and proteomic experiments of Battle et al. 2015 [46]. Individual genotype information for 46 individuals in the Battle et al. study is provided by the 1000 Genomes Project. Protein expression change was evaluated in association with both gain of predicted positive uORFs (ATG and CTG) and loss of predicted positive uORFs. Functional annotation clustering of genes associated with variants examined was performed using Database for Annotation, Visualization and Integrated Discovery v.6.8 (DAVID) [47].

Natural variation affecting predicted positive uORFs:

Polymorphisms that affect the start codons of predicted positive uORFs were identified using data from the 1000 Genomes project. The subset of these SNVs that is associated with differential disease susceptibility was identified through search of the NHGRI-EBI GWAS database. Measurement of comparative frequency of mutation among uORF start codons was taken as a measure of evolutionary conservation and functional significance of predicted positive uORFs.

Cancer mutations affecting predicted positive uORFs:

The study of Alexandrov et al. 2012 [39] provides a set of exomic somatic mutations according to patient sample and cancer type. We used these mutations as a comparison standard for the healthy 1000 Genomes Project population. We identified start codons of our predicted positive uORFs altered by somatic mutation in cancer.

Results:

Genome-wide search yielded 1,270,265 unique uORFs. Within this large set, we isolated the subset of uORFs identified as translated in the studies of Lee et al. 2012, Fritsch et al. 2012, and Gao et al. 2014. We further stratified this set of translated uORFs according to shared representation of uORFs among the three studies. uORFs identified in the intersection between two or more of these studies were used as the reference standard for functional uORFs. Literature review yielded 33 additional examples of active uORFs that were also included in the set of positive, functional uORFs.

We followed the procedure outlined in Fig. 2a to isolate uORFs that are likely to be active. Distributions of attributes for positive, translated uORFs were compared with distributions of those same attributes observed in the set of unlabeled uORFs [Fig. 2b]. The KS statistic and corresponding p-value for each of the 89 attributes assessed in this study are provided in *Supplement Table 2*. The top 10 attributes listed according to magnitude of KS statistic are given in Fig. 2c From this prioritization of attributes, we can draw insights into the relationship between uORF structure and function. The presence of large numbers of start codons within a single uORF is a high priority attribute for positive classification, as is a shorter distance between the uORF and the CDS. ATG is the start codon associated with greatest functional significance. Start and stop codons of functional uORFs are generally located in evolutionarily conserved sites suggesting a meaningful physiologic role.

Overlap between the three ribosome profiling experiments was found to be low, with pairwise intersections of 12.2% (Gao ∩ Fritsch), 9.2% (Gao ∩ Lee), and 9.8% (Lee ∩ Fritsch). The number of uORFs shared between all three sets represents only 3.3% of uORFs identified in these studies [Fig. 3a]. If independent ribosome profiling experiments represent resampling of the same population, repeat identification of uORFs among experiments yields an estimate of the total number of functional uORFs. 10,000 functional uORFs are estimated in this way to be present in the human genome using the Schnabel equation (Eq. 1) or Schumacher and Eschmeyer equation (Eq. 2) [48,49].

CTG (28.2%) and ATG (46.1%) are the most prevalent start codons identified in ribosome profiling experiments. CTG (30.5%) and ATG (34.6%) continue to represent the majority of start codons in intersection between ribosome profiling experiments [Fig. 3b]. Representation of every near-cognate start codon was found in intersections between studies, with the exception of AAG and AGG. This indicates that uORFs do not generally employ AAG and AGG as start codons. Therefore, identification of uORFs beginning with AAG or AGG in ribosome profiling experiments may represent false-positives.

Discretized attributes of positive and unlabeled sets of uORFs were used to build a statistical classifier within a Naive-Bayes framework. The result of application of the classifier is shown in Fig. 3.C. 76.8% of 2-voted positive uORFs [590/768], 67.1% of 1-voted positive uORFs [2,379/3,543], and 14.7% of unlabeled uORFs [185,833/1,265,954] are ultimately classified as likely active. A total of 14.9% of all uORFs are identified as likely active [188,802/1,270,265]. A complete list of upstream open reading frames predicted to be active is provided as *Supplemental Table 5.* The 10% highest probability examples are also specified (Supplemental Table 6).

A large proportion of uORFs in the human genome begin with CTG start codons (19.3%). The greatest number of predicted positive uORFs are also initiated with a CTG start codon (11.8%). ATG has a lower comparative prevalence in the human genome and in the predicted positive set (6.7% and 8.2% respectively) [Fig. 3d]. 8 genes are associated with greater than 200 positively scored uORFs (FAM156B, FAM156A, EEF1D, UBA1, C6orf62, HMGB1, HP1BP3, TBC1D5), suggesting that these genes are under strong and redundant translational regulation mediated by uORFs. The proportion of uORFs ultimately identified as positive from each ribosome profiling study is shown in Fig. 3e. The results were similar for each of the ribosome profiling experiments, approximately 70% in each case (72% of Gao, 71% of Lee, 70% of Fritsch).

As a validation of our technique, we serially excluded one of three ribosome profiling experiments from the positive training set, instead including the excluded set among unlabeled examples for subsequent retrieval [Fig. 3f]. The AUC for each of the ROC curves corresponding to these trials is similar: 0.82, 0.79, and 0.77. Given the low overlap observed between ribosome profiling experiments, this suggests a high false-negative rate for ribosome profiling studies; we believe predicted active uORFs reflect those uORFs that additional experiments would discover are translated.

As experimental validation of our technique, we examined how natural variation affecting our predicted active uORFs alters protein level and ribosome localization in humans. We hypothesized that an active uORF altered by naturally occurring variants should create observable effect on ribosome occupancy and protein levels from that gene. The results of Battle et al. 2015, supplemented by genotype information from the 1000 Genomes Project, provide the basis for validation of our predictions in 46 human subjects (*Supplemental Table 7*).

In this study of natural variation amongst humans, variants causing gain of predicted positive ATG or CTG uORFs are associated with increase in downstream protein expression. Variants that cause loss of predicted positive uORFs are associated with decrease in downstream protein expression [Fig. 4a]. That is, there is a statistically significant difference in mean protein expression between variants causing uORF gain compared with uORF loss, among variants with approximate balance between individuals with and without the variant (N_{loss}=133, $N_{gain=}$17, t= 2.6, DOF=307, p=0.011, for variants shared by >10 individuals). A case example, documenting possible effect related to uORF gain associated with the gene EIF5A, is documented in *Supplemental Results* and *Supplemental Fig. S.1*.

We hypothesized the observation of decreased protein level with uORF loss may relate to uORF-uORF repression: a uORF upstream of another uORF may repress the downstream uORF. Upon restricting our protein level analysis to uORFs least likely to repress downstream uORFs – uORFs directly overlapping the CDS – we observed a trend towards increases in protein levels with uORF loss ($N_{CDSoverlap}$ = 34, μ=0.065; $N_{CDSnon\text{-}overlap}$ = 99, μ=-0.055; p=0.097, for variants shared by > 10 individuals). This is consistent with the classical role of uORFs as translational repressors. We also considered that observed protein level changes might in some cases relate to multiple uORFs affected by a single variant. Among predicted positive uORFs affected by start codon altering variants, 3.6% of these variants caused simultaneous truncation of an overlapping predicted positive uORF.

Functional annotation clustering of genes associated with variants affecting predicted positive uORFs, showed greatest enrichment for ribosomal proteins including RPL24 (32 associated with uORF loss and 17 associated with uORF gain), and ribosome associated proteins including EIF3 (DAVID enrichment score 20.94, N_{terms}=12, $N_{genes(enrich.)}$/$N_{genes(tot.)}$=108/961, $p_{mean(geom.)}$<<0.001). EIF3 and ribosomal proteins like RPL24 are thought to overcome uORF mediated repression in *Arabidopsis thaliana* through facilitation of translation reinitiation [50].

For these same 46 human subjects, cis-rQTLs provide an inventory of variants with statistically significant effect on local ribosome occupancy. There is significant enrichment for rQTLs interrupting positively scored start codons [Fig. 4b]. If mutations hit uORFs randomly, 14.9% of the time they would hit a positively scored uORF. However, we observe that 48% of these rQTLs (21/44) interrupt positively scored start codons -- a 3x higher rate. This indicates that many rQTLs may measure the direct effect of disruption of functional uORFs.

The ATG start codon is relatively conserved among predicted positive start codons -- it is rarely interrupted by 1000 Genomes Project variants (relative rate (RR) 0.03), suggesting its functional importance. The CTG start codon, although more prevalent among predicted positive uORFs, is altered relatively frequently by natural human variants (RR 0.52) [Fig. 4c]. In exomic tumor samples from cancer patients, CTG is the most commonly modified predicted positive uORF start codon. ATG is interrupted at a RR of 0.25 in comparison to CTG [Fig. 4d]. The higher RR of interruption of both ATG and CTG in cancer as compared to germline variants – 8 fold higher, and 2 fold higher respectively – further suggests functional consequences attributable to these uORFs.

Exomic cancer mutations breaking the highest scored uORFs are listed in *Supplemental Table 8*. These mutations interrupt uORFs associated with well-studied oncogenes and tumor suppressors. MYC and BCL2 are the two genes associated with the greatest recurrence of uORF interruptions, and we identify recurrent mutation of positively scored uORFs associated with PTEN, TP53, ERCC1, and MSH5. Genome-wide association study (GWAS) SNVs listed in the NHGRI-EBI GWAS database that impact our predicted uORFs are listed in *Supplemental Table 4*. GWAS diseases associated with SNVs interrupting positively scored uORFs include prevalent chronic conditions like obesity (rs11603334), osteoporosis (rs3755955), asthma (rs3771180), and type 2 diabetes (rs1552224). Additional variants associated with susceptibility and prognosis in cancer are found to interrupt positively scored uORFs, like rs779805 upstream of the VHL gene, and rs34330 upstream of CDKN1B. Although linkage disequilibrium and overlap among regulatory elements complicates interpretation of these GWAS studies, these disease associated SNVs may owe their functional consequence to alteration of a translated uORF.

Discussion:

In this study, we identify 188,802 likely active upstream open reading frames from a genome-wide set of 1,270,265 unique uORFs. We further highlight the 10% of our predictions that are most likely to be functional, as a high reliability subset.

We began by assuming that ribosome profiling experiments have a high false negative rate for identification of functional uORFs. Our method applied the intersection of three ribosome profiling studies, to form a reference set of known active uORFs. The low overlap between ribosome profiling experiments suggests a high false-negative rate in individual experiments. The finding that pairs of ribosome profiling experiments may be used to correctly identify the uORFs translated in a third experiment also suggests a high false negative rate. The large number of uORFs we identified as likely functional is consistent with this premise, but significant in comparison to other studies on the topic.

There is precedent for our findings, in comparisons of large-scale parallel experiments of interaction between biomolecules. The protein-protein interaction experiments of Uetz et al. employed a comprehensive, genome-wide scope [51]. Subsequent experiments by Ito et al., with similar technique and scope, showed low overlap with results of the prior project [52]. It became clear that the universe of possible protein-protein interactions is much larger than identified in either experiment individually. Benefit in identifying these interactions is achieved by combining datasets [53].

Our use of an intersection between ribosome profiling experiments provides some control against differences in experimental conditions and tissue specific results (both HEK293 and THP-1 cells were examined). However, just as protein levels vary widely across cell-types [54], it may prove that the activity of uORFs varies considerably across cell types and cellular conditions. Analysis of cell-type specific and condition specific activity of uORFs may further expand estimates of the population of translated uORFs.

Our study helps clarify how attributes of structure and context of a given uORF -- including start codon, base composition, and relative position to the CDS -- likely contribute to varying functionality among uORFs. Although ATG is the most common uORF start codon identified in ribosome profiling experiments, lower affinity near cognate-start codons may have great functional impact on the landscape of translation, due to their overall abundance.

An important validation of our predictions is the finding that alteration of predicted functional uORFs, as a consequence of germline genetic variation, impacts ribosome binding and protein level in humans. Generally we assume that uORFs act as translational repressors. However, the overall effect of uORF loss appears to be a decrease in downstream protein level. This is contrary to common view that uORFs act as translational repressors. Mechanisms have been studied, where uORFs act to up-regulate expression of a downstream coding sequence (e.g. leaky-scanning, and translation reinitiation). Ribosomal reinitiation at an uORF on the ATF4 gene, is one particularly well studied example of such a mechanism [55]. Our analysis suggests that positive effect on translation may be a more common consequence for upstream open reading frames than is previously credited.

Reasons for our observed protein level changes may also include multiple indirect effects of uORF repression such as A) uORF-uORF interaction where one uORF acts to repress another uORF, B) variation affecting overlapping uORFs simultaneously, and C) uORFs upstream of coding genes that themselves regulate translation. Indeed, the observation of enrichment of translational mediators and ribosomal proteins among our uORFs affected by genetic variation, suggests the possibility of cascading functional effects related to uORF gain or loss. Furthermore, among genes with multiple predicted positive uORFs, the presence of CDS-overlapping uORFs resulted in opposite effect on CDS translation compared to those uORFs entirely upstream of the CDSs. This observation suggests that the effect of interaction among uORFs is worthy of further study.

Our results suggest avenues for future research. Identification of human germline variants altering predicted positive uORFs reveals locations where the creation or destruction of an uORF is likely to alter protein levels. Employing this method, we identified disease associated SNVs -- including a number of GWAS SNVs -- that likely owe their significance to alteration of a functional uORF. Among diseases, our work could be used to help broaden knowledge of the role of uORFs in cancer beyond recently identified individual examples [56].

Finally, we provide a catalog that can serve as a point of reference for other researchers engaged in the investigation of uORF function.

References:

1. Kozak, M. An analysis of 5’-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.* **15,** 8125–8148 (1987).

2. Kochetov, A. V., Sarai, A., Rogozin, I. B., Shumny, V. K. & Kolchanov, N. A. The role of

alternative translation start sites in the generation of human protein diversity. *Mol. Genet. Genomics* **273,** 491–496 (2005).

3. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* **147,** 789–802 (2011).

4. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–23 (2009).

5. Ivanov, I. P., Loughran, G. & Atkins, J. F. uORFs with unusual translational start codons autoregulate expression of eukaryotic ornithine decarboxylase homologs. *Proc. Natl. Acad. Sci.* **105,** 10079–10084 (2008).

6. Ivanov, I. P., Firth, A. E., Michel, A. M., Atkins, J. F. & Baranov, P. V. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Res.* **39,** 4220–4234 (2011).

7. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci.* **106,** 7507–7512 (2009).

8. Johnstone, T. G. *et al.* Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* **35,** 706–723 (2016).

9. Somers, J., Pöyry, T. & Willis, A. E. A perspective on mammalian upstream open reading frame function. *Int. J. Biochem. Cell Biol.* **45,** 1690–1700 (2013).

10. Meijer, H. A. & Thomas, A. A. M. Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA. *Biochem. J.* **367,** 1–11 (2002).

11. Barbosa, C. *et al.* Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLoS Genet.* **9,** e1003529 (2013).

12. Morris, D. R. & Geballe, A. P. Upstream Open Reading Frames as Regulators of mRNA Translation. *Mol. Cell. Biol.* **20,** 8635–8642 (2000).

13. Sonenberg, N. & Hinnebusch, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136,** 731–45 (2009).

14. Hinnebusch, A. G. *et al.* Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352,** 1413–6 (2016).

15. Chua, J. J. E. *et al.* Synthesis of two SAPAP3 isoforms from a single mRNA is mediated via alternative translational initiation. *Sci. Rep.* **2,** 277–298 (2012).

16. Mackowiak, S. D. *et al.* Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* **16,** 179 (2015).

17. Andrews, S. J. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* **15,** 193–204 (2014).

18. Oyama, M. *et al.* Analysis of Small Human Proteins Reveals the Translation of Upstream Open Reading Frames of mRNAs. *Genome Res.* **14,** 2048–2052 (2004).

19. Bergeron, D. *et al.* An Out-of-frame Overlapping Reading Frame in the Ataxin-1 Coding Sequence Encodes a Novel Ataxin-1 Interacting Protein. *J. Biol. Chem.* **288,** 21824–21835 (2013).

20. Starck, S. R. *et al.* Translation from the 5' untranslated region shapes the integrated stress response. *Science* **351,** aad3867 (2016).

21. Andreev, D. E. *et al.* Oxygen and glucose deprivation induces widespread alterations in mRNA translation within 20 minutes. *Genome Biol.* **16,** 90 (2015).

22. Shalgi, R. *et al.* Widespread Regulation of Translation by Elongation Pausing in Heat Shock. *Mol. Cell* **49,** 439–452 (2013).

23. Wiita, A. P. *et al.* Global cellular response to chemotherapy-induced apoptosis. *Elife* **2,** e01236 (2013).

24. Gerashchenko, M. V., Lobanov, A. V. & Gladyshev, V. N. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc. Natl. Acad. Sci.* **109,** 17394–17399 (2012).

25. Liu, B., Han, Y. & Qian, S.-B. Cotranslational Response to Proteotoxic Stress by Elongation Pausing of Ribosomes. *Mol. Cell* **49,** 453–463 (2013).

26. Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44,** 283–292 (1986).

27. Brar, G. a & Weissman, J. S. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* **16,** 651–664 (2015).

28. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15,** 205–213 (2014).

29. Gao, X. *et al.* Quantitative profiling of initiating ribosomes in vivo. *Nat. Methods* **12,** 147–153 (2014).

30. Fritsch, C. *et al.* Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.* **22,** 2208–2218 (2012).

31. Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci.* **109,** E2424–E2432 (2012).

32. Selpi, S. *et al.* Predicting functional upstream open reading frames in Saccharomyces cerevisiae. *BMC Bioinformatics* **10,** 451 (2009).

33. Hu, Q., Merchante, C., Stepanova, A. N., Alonso, J. M. & Heber, S. Genome-Wide Search for Translated Upstream Open Reading Frames in Arabidopsis Thaliana. *IEEE Trans.*

Nanobioscience **15,** 148–157 (2016).

34. Fields, A. P. *et al.* A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol. Cell* **60,** 816–827 (2015).

35. Raj, A. *et al.* Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* **5,** (2016).

36. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).

37. McVean, G. A. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

38. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, and P. H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **Vol. 42,** D1001–D1006 (2014).

39. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–21 (2013).

40. Wen, Y. *et al.* Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nat. Genet.* **41,** 228–233 (2009).

41. Raveh-Amit, H. *et al.* Translational Control of Protein Kinase C by Two Upstream Open Reading Frames. *Mol. Cell. Biol.* **29,** 6140–6148 (2009).

42. Fayyad, U. & Irani, K. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. *donga.ac.kr*

43. Liu, B., Dai, Y., Li, X., Lee, W. S. & Yu, P. S. Building text classifiers using positive and unlabeled examples. in *Third IEEE International Conference on Data Mining* 179–186 (IEEE Comput. Soc, 2003). doi:10.1109/ICDM.2003.1250918

44. Rish, I. An empirical study of the naive Bayes classifier. *researchgate.net*

45. Zhang, H. The Optimality of Naive Bayes. *AA* (2004).

46. Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347,** 664–7 (2015).

47. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2008).

48. Schnabel, Z. E. The Estimation of Total Fish Population of a Lake. *Am. Math. Mon.* **45,** 348 (1938).

49. Schumacher, F. X. and Eschmeyer, R. W. The estimation of fish populations in lakes and ponds. *J. Tennessee Acad. Sci.* **18,** (1943).

50. Xue, S. & Barna, M. Specialized ribosomes: a new frontier in gene regulation and

organismal biology. *Nat. Rev. Mol. Cell Biol.* **13,** 355–369 (2012).

51. Fields, S. *et al.* A comprehensive analysis of protein|[ndash]|protein interactions in Saccharomyces cerevisiae. *Nature* **403,** 623–627 (2000).

52. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 4569–74 (2001).

53. Jansen, R. & Gerstein, M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. Microbiol.* **7,** 535–545 (2004).

54. Pontén, F. *et al.* A global view of protein expression in human cells, tissues, and organs. *Mol. Syst. Biol.* **5,** 799–816 (2009).

55. Vattem, K. M. & Wek, R. C. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 11269–74 (2004).

56. Wethmar, K. *et al.* Comprehensive translational control of tyrosine kinase expression by upstream open reading frames. *Oncogene* **35,** 1736–1742 (2016).

Acknowledgments:

This work was supported by the Yale University School of Medicine Medical Student Research Fellowship, and Leon Rosenberg, M.D. Medical Student Research Fund in Genetics and Jane Danowski Weiss Family Foundation Fellowship at Yale University School of Medicine to P.D.M.

Author contributions:

S.B. designed the study. S.B., M.G. and R.K. supervised the study. R.A., P.M. and M.P. wrote analysis software. P.M. and R.A. analyzed study data. P.M. and M.G. wrote the paper. S.B., R.K. and R.A. edited the paper.

Competing financial interests:

The authors have no competing financial interests to declare.

Figures:

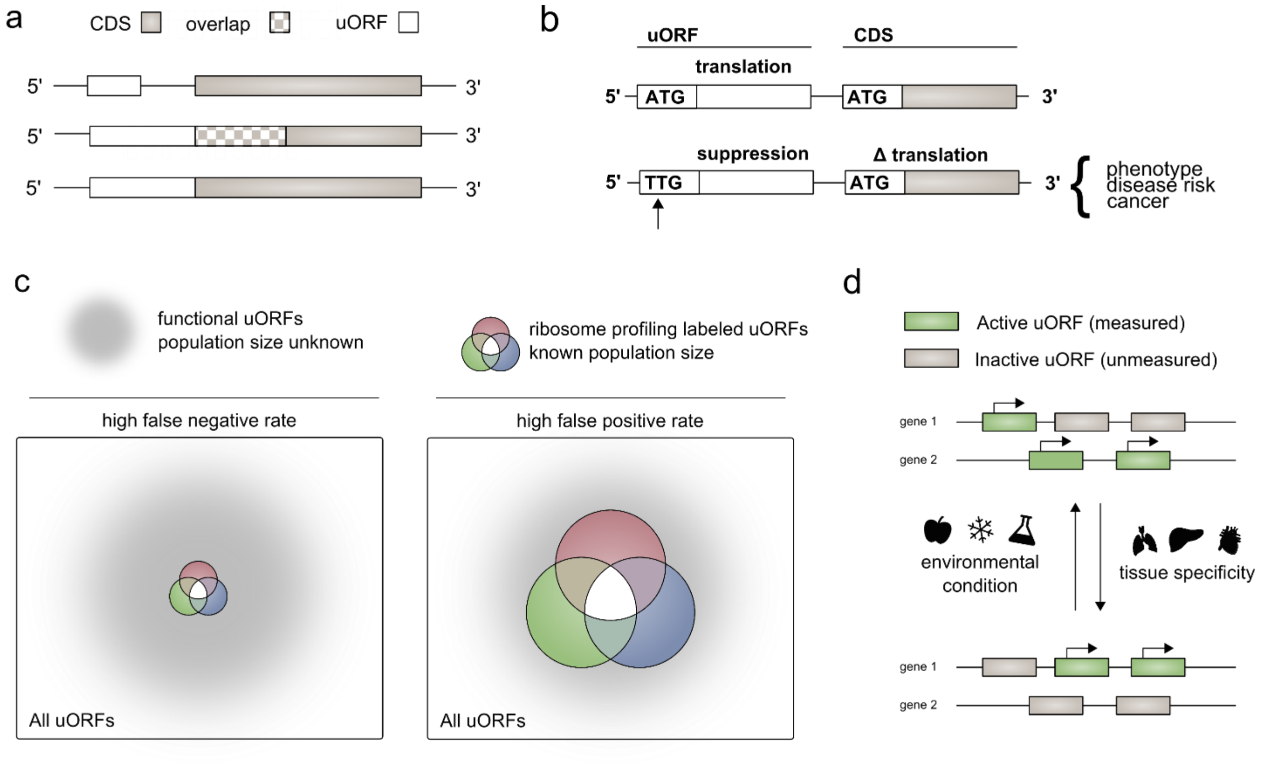

Figure 1:

(a) Structure of upstream open reading frames. The stop codon of an uORF may be located before the CDS start codon [top], or downstream of the CDS start codon, if the uORF is frame-shifted relative to the CDS [middle]. If the uORF and CDS share the same stop codon, the uORF acts as a 5' extension of the CDS [bottom]. ***(b) Effect of mutation or variation on upstream open reading frames.*** Creation or destruction of an upstream open reading may have downstream effect on translation of the coding sequence. Change in translation of the coding sequence may result in change in phenotype and disease risk. ***(c) Sensitivity and specificity of ribosome profiling for identifying upstream open reading frames.*** It is possible that ribosome profiling studies have a high false negative rate (left), or a high false positive rate (right). We make the assumption that ribosome profiling studies have a high false negative rate for identifying translated upstream open reading frames (left). ***(d) Activity of uORFs varies according to cell type and environmental stimuli.*** uORFs may not be detected in a ribosome profiling experiment due to variation in uORF activity with cell type and cell environment.

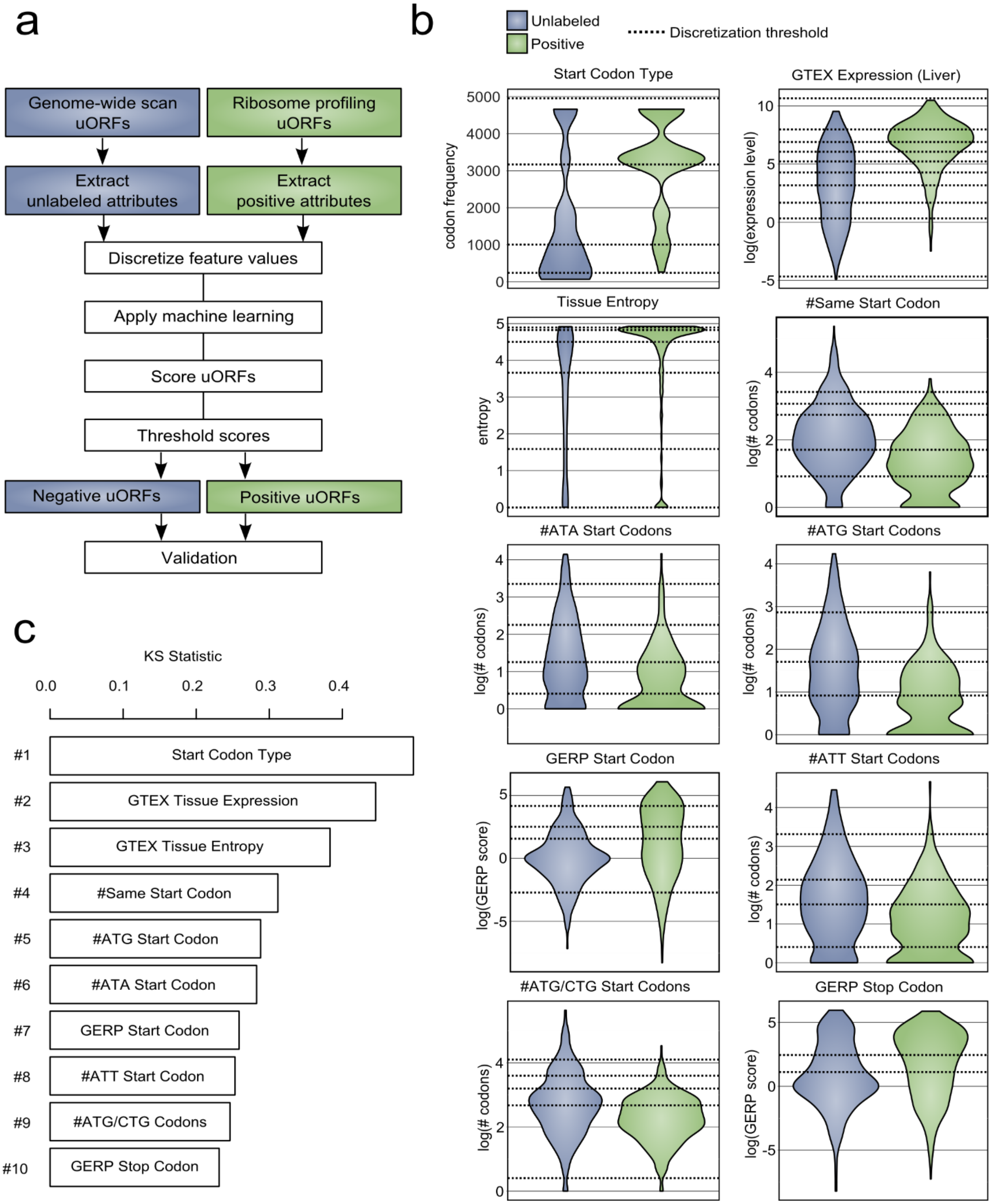
a
Genome-wide scan uORFs
Ribosome profiling uORFs
Extract unlabeled attributes
Extract positive attributes
Discretize feature values
Apply machine learning
Score uORFs
Threshold scores
Negative uORFs
Positive uORFs
Validation
b
Unlabeled
Positive
Discretization threshold
Start Codon Type
codon frequency
5000
4000
3000
2000
1000
0
GTEX Expression (Liver)
log(expression level)
10
5
0
-5
Tissue Entropy
entropy
5
4
3
2
1
0
#Same Start Codon
log(# codons)
4
2
0
#ATA Start Codons
log(# codons)
4
3
2
1
0
#ATG Start Codons
log(# codons)
4
3
2
1
0
GERP Start Codon
log(GERP score)
5
0
-5
#ATT Start Codons
log(# codons)
4
3
2
1
0
#ATG/CTG Start Codons
log(# codons)
4
2
0
GERP Stop Codon
log(GERP score)
5
0
-5
c
KS Statistic
0.0
0.1
0.2
0.3
0.4
#1 Start Codon Type
#2 GTEX Tissue Expression
#3 GTEX Tissue Entropy
#4 #Same Start Codon
#5 #ATG Start Codon
#6 #ATA Start Codon
#7 GERP Start Codon
#8 #ATT Start Codon
#9 #ATG/CTG Codons
#10 GERP Stop Codon

Figure 2:

(a) Methodology for distinguishing positive from unlabeled uORFs. uORFs identified through genome-wide scan and uORFs labeled in ribosome profiling experiments were used to train a machine learning algorithm to identify uORFs that are likely active (positive predictions). ***(b) Distributions of attributes for positive and unlabeled uORFs.*** uORF attributes are used to distinguish positive from unlabeled uORFs. Continuous distributions were discretized and optimized for machine learning using the minimum description length principle (MDLP) binning algorithm. Horizontal lines on the plot correspond to these binning intervals. The 10 attributes with the greatest difference in distribution (largest Kolmogorov Smirnov (KS) statistic) between positive and unlabeled uORFs are shown. ***(c) Upstream open reading frame attributes as classifiers.*** Attributes are ranked according to the difference in distribution between positive and unlabeled uORFs, using the KS statistic.

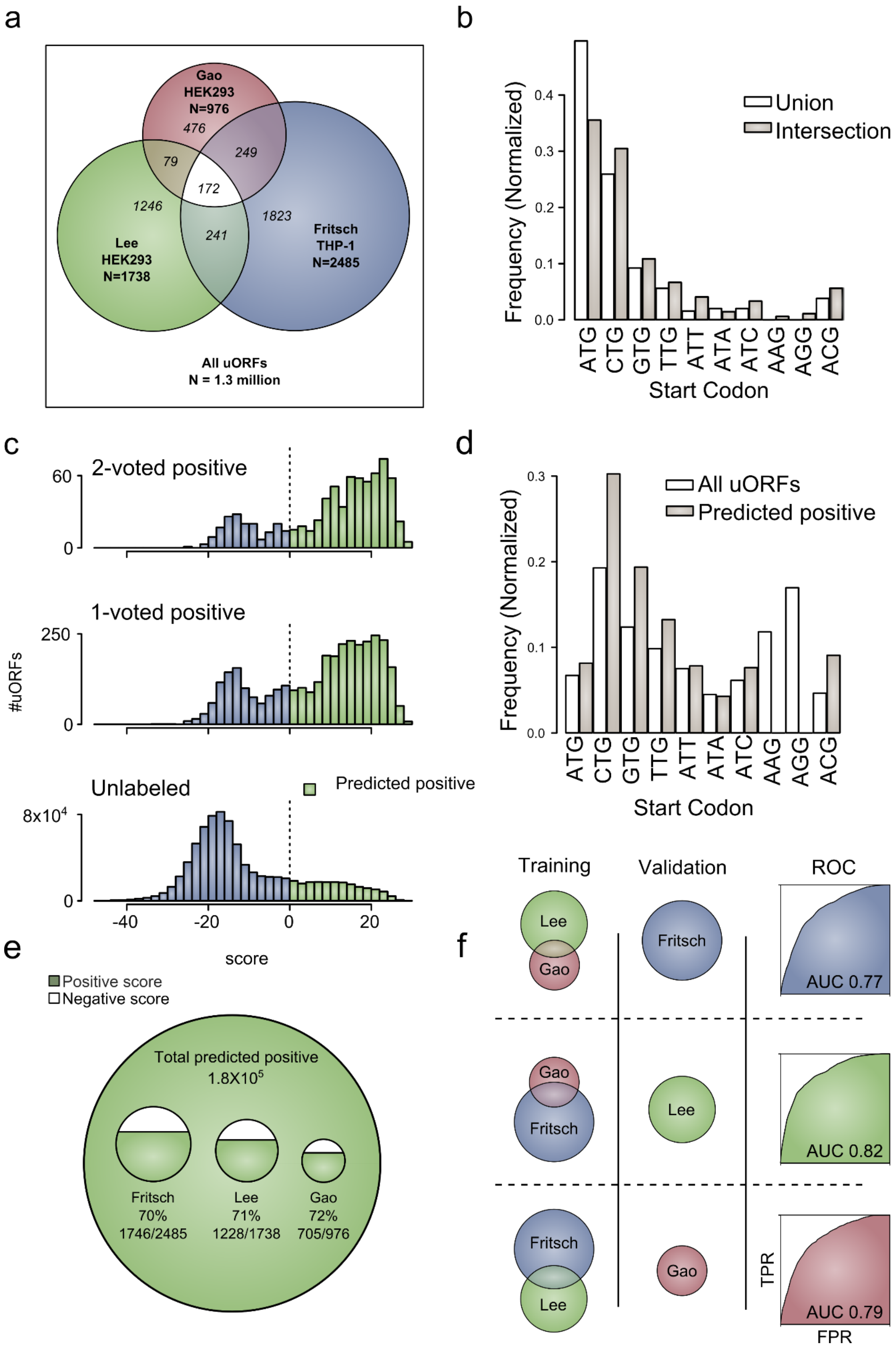
a
Gao
HEK293
N=976
476
249
79
172
1246
241
1823
Lee
HEK293
N=1738
Fritsch
THP-1
N=2485
All uORFs
N = 1.3 million
b
Frequency (Normalized)
0.0
0.1
0.2
0.3
0.4
Union
Intersection
ATG
CTG
GTG
TTG
ATT
ATA
ATC
AAG
AGG
ACG
Start Codon
c
2-voted positive
60
0
1-voted positive
250
0
#uORFs
Unlabeled
Predicted positive
$8x10^4$
0
-40
-20
0
20
score
d
Frequency (Normalized)
0.0
0.1
0.2
0.3
All uORFs
Predicted positive
ATG
CTG
GTG
TTG
ATT
ATA
ATC
AAG
AGG
ACG
Start Codon
e
Positive score
Negative score
Total predicted positive
$1.8X10^5$
Fritsch
70%
1746/2485
Lee
71%
1228/1738
Gao
72%
705/976
f
Training
Validation
ROC
Lee
Gao
Fritsch
AUC 0.77
Gao
Fritsch
Lee
AUC 0.82
Fritsch
Lee
Gao
AUC 0.79
TPR
FPR

Figure 3:

(a) Frequency of translated uORF ATG start codons and near-cognate start codons, from ribosome profiling experiments. Frequency for uORFs translated in any experiment (union), or in more than one experiment (intersection). ***(b) Ribosome profiling identified uORFs as a subset of all uORFs.*** The universe of all uORFs is identified through comprehensive search of the GENCODE human genome annotation [outer border]. Ribosome profiling studies of Fritsch et al., Lee et al., and Gao et al. are shown as overlapping subsets of this universe. Pair-wise and three-way intersections between these experiments are highlighted. ***(c) Score distributions for upstream open reading frames.*** Score distributions for 2-voted positive uORFs that are translated in two or more ribosome profiling experiments (top), 1-voted positive uORFs that are translated in only one ribosome profiling experiment (middle), and unlabeled uORFs uncovered through genome-wide search (bottom). ***(d) The frequency of uORF ATG start codons and near-cognate start codons, for predicted positive upstream open reading frames.*** Frequency is given for all uORFs genome-wide, and for the subset of uORFs that are predicted to be active (predicted positive). ***(e) uORFs predicted as positive from genome-wide scan and ribosome profiling experiments.*** Approximately 180 000 uORFs in the genome are predicted as active upstream open reading frames. This large set includes substantial proportions uORFs identified in the ribosome profiling experiments (~70% each). ***(f) Performance of the machine learning algorithm.*** The machine learning algorithm was trained on two of three ribosome profiling data sets and used to extract the third data set from among unlabeled examples. The ROC curve is shown for each of the three combinations: 1) Train Lee et al. and Fritsch et al. – extract Gao et al. (AUC = 0.79), 2) Train Lee et al. and Gao et al. – extract Fritsch et al. (AUC = 0.77). 3) Train Fristch et al. and Gao et al. - extract Lee et al. (AUC = 0.82).

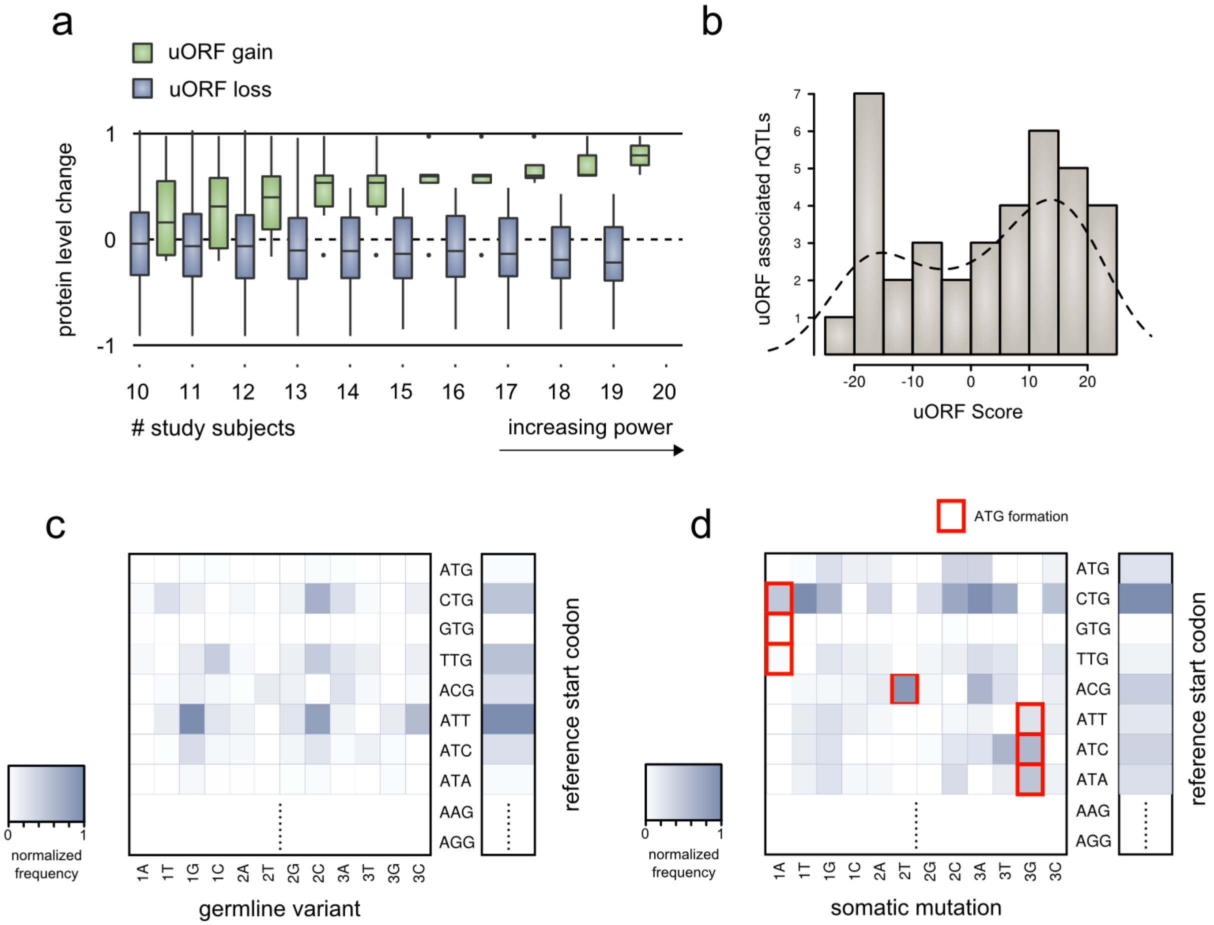

Figure 4:

(a) Gene level protein expression change between individuals with variants interrupting predicted positive uORFs and wild type individuals. uORF gain is associated with increased protein expression, while uORF loss is associated with decreased protein expression. ***(b) rQTLs interrupting uORFs, according to score of the corresponding uORF.*** rQTLs show bias towards interrupting positively predicted uORFs. ***(c) Density matrix showing the distribution of 1000 Genomes variants that interrupt predicted positive uORF start codons.*** The vertical axis displays the reference start codon, and the horizontal axis shows the interrupting variant (position – 1,2,3 – and codon – A,T,G,C). ***(d) Density matrix showing the distribution of somatic mutations found in exomic tumor samples that interrupt predicted positive uORF start codons.*** The vertical axis displays the reference start codon, the horizontal axis shows the interrupting variant (position – 1,2,3 – and codon – A,T,G,C). ATG forming mutations are highlighted.